Analyzing real life loan default data set:

Data cleaning issues:

1. There were a few columns in the data set that has more than 70% of the data missing.

I deleted those columns.

1. There were columns that had date information in the “DEC-2011” string format. I didn’t know how to convert them to Date format in R.

Moreover, I didn’t know how dates would be useful for the model in predicting defaults. So, I removed those columns.

1. Certain cases had ambiguous loan status information like “Does not meet the credit policy. Status:Current”.

Since the meaning of this status was not clear (as in this loan was a good one or a bad one), I removed such cases from the dataset.

1. Some columns had categorical data in string format.

I converted them to factors.

Issues encountered while training the model:

1. The first time I ran the command to build the model using logistic regression in R, I got the following message:

*Warning messages:*

*1: glm.fit: algorithm did not converge*

*2: glm.fit: fitted probabilities numerically 0 or 1 occurred*

1. Out of curiosity, I wanted to see if my dataset had perfect multicollinearity issues. I ran the command to create a correlation matrix for all the continuous predictors in the data set.

I got the following error message:

*Warning message:*

*In cor(loanData.clean[, c("loan\_amnt", "funded\_amnt", "funded\_amnt\_inv", :*

*the standard deviation is zero.*

1. Next, I decided to build the model using forward step wise selection

manually. I did this with the idea that I would be able to find out exactly which predictor variable was causing my model to “break”.

After doing forward selection, I was able to find which predictors were causing the model to break. If I excluded these predictors from the set of inputs to the model, R was able to give create a logistic regression model from the training set.

1. After obtaining the logistic regression the model, I wanted to check for perfect multicollinearity problems in the model again.

I ran the VIF(model) command and I got the following message:

*Error in vif.default(loanData.logModel) : there are aliased coefficients in the model.*

After “googling” this error message, I learned that this error is usually caused due to perfect multicollinearity issues in the dataset.

1. To investigate this perfect multicollinearity problem, I tried getting the correlation matrix again for all the continuous predictors that I gave as input to the model that I just built.

I got the following message:

Warning message:

In cor(loanData.clean[, c("loan\_amnt", "int\_rate", "installment", : the

standard deviation is zero.

1. To investigate the standard deviation is zero problem, I ran commands to find out the standard deviations of all predictors one by one. There was one column that had all zeroes. This one was the culprit. I removed it from the dataset. Now, I was able to obtain the correlation matrix.
2. In the correlation matrix, I found that there were some predictors that were highly correlated. First, I removed those predictors from the training dataset and built the model. The residuals were very high.

I rebuilt the model now including those predictors that I had excluded earlier. The residuals are much lower.

1. When I ran the predict() command, I got the following message:

Warning message:

In predict.lm(object, newdata, se.fit, scale = 1, type = ifelse(type == : prediction from a rank-deficient fit may be misleading.

1. But, I still got the probabilities for each case and I was able to classify each case as a good loan or bad loan using a probability threshold.
2. I was able to predict the results in the test set and I achieved almost 100% accuracy. This bothers me. I think I made a mistake somewhere.